主动地纠错式半监督聚类社区发现算法 *

张贤坤，刘渊博†，任　静，张高祯

(天津科技大学 计算机科学与信息工程学院，天津 300457)

摘　要：经典的无监督聚类算法快速、简单且可以直接对大规模数据集进行划分，但是由于网络结构较为复杂，划分的准确度并不高。为此，提出一种基于主动学习的纠错式半监督社区发现算法 ESCD（error correction semi-supervised community detection algorithm），将传统的 K-means 算法进行分步计算，并且在聚类的过程中加入成对约束。根据先验信息保留正确的划分，纠正错误的划分来改变网络的连接关系，使网络具有更明显的块结构，当节点与聚类中心的距离不再变化时划分结束。实验结果表明，与现有的社区发现算法相比，ESCD 算法具有更高的精度，且所需的监督信息远远小于其他半监督算法。

关键词：主动学习；纠错式半监督社区发现；K-means 算法；成对约束

中图分类号：TP301.6　　doi: 10.3969/j.issn.1001-3695.2018.03.0162

Active error-correcting community discovery algorithm based on semi-supervised clustering

Zhang Xiankun, Liu Yuanbo†, Ren Jing, Zhang Gaozhen

(*School of Computer Science & Information Engineering*, *Tianjin University of Science & Technology*, *Tianjin* 300457, *China*)

Abstract: The classical unsupervised clustering algorithm is fast, simple and suitable for mining large-scale datasets, and it can also directly divide communities. However, due to the complexity of communities, the classification accuracy of the algorithm is not ideal. Therefore, this paper proposes an error-correcting semi-supervised community detection algorithm (ESCD) based on active learning. It can calculate the traditional k-means algorithm step by step, and adding pairs of constraints in the clustering process. In order to preserve the correct partitioning according to the prior information, we correct the wrong division to change the connection of the network. So that the network has a more obvious block structure in the process of changing the distance between nodes and cluster centers. The results of the experiment show that compared with the existing community discovery algorithms, the ESCD algorithm has higher accuracy with less supervisory information than other semi-supervised algorithms.

Key words: active learning; error correction semi-supervised community discovery; K-means algorithm; constraints in pairs

0　引言

近年来，随着对复杂网络研究的深入，研究者们发现很多实际网络都是由社区构成的，复杂网络内部连接紧密的节点组成的集合就称之为社区，同一社区内部的节点之间联系较为密切，而不同社区之间的联系较为松散[1]。社区发现是很多领域的研究热点，因为社区发现可以更准确地定位社会群体，有助于将具有相似兴趣的相关人员联系起来，以便与有共同兴趣的群体分享他们的想法和专业知识[2]。

复杂网络聚类本质上是图的划分问题[3]，因此使用传统聚类算法对复杂网络进行聚类在社区发现领域得到了广泛应用，例如，K-means、K-mediods、谱聚类和图聚类等。其中 K-means 算法[4]是 MacQueen 于 1967 年提出的一种经典的聚类算法，其特点高效简洁，被广泛应用于数据聚类中，在网络数据的处理中也得到成功应用。该算法的核心思想为找出 K 个聚类中心，使得每个数据点和与其最近的聚类中心的平方距离和被最小化。然而，真实的网络结构的复杂度较高，社区属性模糊，无监督聚类不包含任何的先验信息，划分结果仅根据节点特征向量的距离来计算，噪点或不规则的拓扑结构对节点的类归属影响很大，因此完全不包含先验信息的聚类算法稳定度较差；真实网络通常结构较为复杂，社区内部连接不够紧密而社区之间的连接却很多，不清晰的网络结构导致部分位于社区边界的节点无法正确的划分。不包含任何先验信息的社区发现结果依赖于初始聚类中心的选取且当网络结构较复杂时划分准确度会大幅度

收稿日期：2018-03-16；**修回日期：**2018-04-24　**基金项目：**国家自然科学基金资助项目（61702367）；天津市教委科研计划资助项目（2017KJ033）

作者简介：张贤坤，男，安徽芜湖人，教授，博士，主要研究方向为智能信息处理、社会网络分析；刘渊博，女（通信作者），山西太原人，硕士研究生，主要研究方向为智能信息处理（yuanbo1007@mail.tust.edu.cn）；任静，女，山西运城人，硕士研究生，主要研究方向为智能信息处理；张高祯，男，河南周口人，硕士研究生，主要研究方向为智能信息处理.

降低，因此部分半监督社区发现算法被提出[5-7]。半监督学习同时使用了标记数据和未标记数据可以大大提升学习的效率[8]，但是现有的半监督算法需要大量的标记数据，而且无法真正的改变复杂网络的结构，仍存在效率较低的缺陷。

根据上述存在的问题，本文提出一种主动地纠错式半监督社区检测算法 ESCD（error correction semi-supervised community detection algorithm）。将主动学习的思想，运用在半监督聚类算法中。本文将传统的 K-means 算法进行分步计算；将每一步的距离迭代结果视为粗聚类结果，并且根据当前粗聚类结果计算节点隶属度；主动的添加少量先验信息，通过加入逻辑推理充分利用先验信息；改变了复杂网络的结构，得到准确的划分结果。在本文提出的 ESCD 算法中，充分利用了 K-means 算法每一步的迭代过程，并且可以在每一步的距离迭代中自动地纠正粗聚类结果中划分错误的节点，使复杂网络具有更加清晰的块结构。实验结果证明 ESCD 算法具有更高的精度，并且大大提升了原有半监督聚类算法的效率。

1　相关工作

1.1　K-means 算法

本文采用的是基于划分的无监督聚类方法中应用最广泛的 K-means 算法[4]，采用距离作为相似性的评价指标。把数据集划分成 K 个互不交叠的类簇，得到类内高度相似，类间相似度低的划分结果。

给定一个网络图 $G=(V,E)$，其中节点的集合为 $V=\{v_1,v_2,\ldots,v_n\}$；链接的集合为 $E=\{e_1,e_2,\ldots,e_n\}$。邻接矩阵 $A=[a_{ij}]_{n\times n}$ 可以直观的反应节点之间的连接关系，如果 v_i 和 v_j 之间存在链接则 $a_{ij}=1$；如果 v_i 和 v_j 之间不存在链接则 $a_{ij}=0$。在聚类的过程中，将邻接矩阵作为算法的输入，n 个节点的邻接矩阵向量作为节点的 n 维特征，根据节点与 k 个聚类中心 C_j 的多次距离迭代，直到划分结果不再改变，得到最终的社区划分结果。距离用式(1)来度量，其中 $d(x_i,C_j)$ 是节点 x_i 和 C_j 之间的欧氏距离。

$$d(x_i,c_j)=\sqrt{\sum_{\alpha=1}^{m}\left(x_{i\alpha}-c_{j\alpha}\right)^2},i=1,2,\ldots,n;\ j=1,2,\ldots,n \quad (1)$$

1.2　半监督学习算法

半监督学习中辨别信息形式有多种，最常见的是样本类标号，即明确指定每个样本的类别。除类标号之外，还有样本之间的成对约束，成对约束是指两个样本之间的一种关系，包括正约数和负约束。两个样本的类标号相同时，他们之间存在一种正约束关系，反之，两个样本之间存在一种负约束关系[8]。相较于类标号，成对约束对先验信息的需求较小，而且更容易获得。例如，Liu 等人[9]利用标签传播方法，将已标注的节点类标号向周围邻居进行传播。Silva 等人[10]基于模块度最大化的准则将半监督方法融入社区发现中。Zhang 等人[11]直接将节点的成对约束加到待分解的邻接矩阵上。Yang 等人[12]利用半监督的潜在空间图正则化方法建立了一个统一的社区检测框架。上述方法存在部分共同的缺陷，就是半监督学习的效率较低，

如图 1 所示，在一次完整的社区发现算法框架之外，以随机的方式将成对约束混入网络中，通过三对成对约束的指导才将节点 5 划分到正确社区内。显然这种随机的标注方式是冗余且低效的，而且框架外的添加模式局限了半监督学习的效率。主要原因如下：a）标记数据添加的位置在完整的算法框架外，一次完整的算法包括多次迭代，这种添加方式没有充分利用每一步的迭代过程；b）先验信息通过随机标注的方式干预社区的划分，需要大量的先验信息，但通常人工标注的先验信息很难获得而且代价昂贵；c）即使加入大量的先验信息依旧无法真正干预网络的正确划分，即随机加入的标签并不是网络划分最需要的那部分；d）成对约束只是指导了部分节点的划分，并没有真正的改变网络的复杂结构，因此无法从根本使网络拥有更明显的块结构。

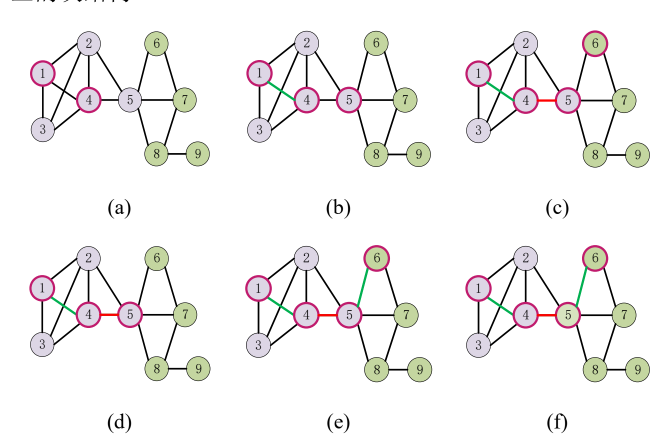


图 1　半监督算法成对约束添加过程

因此本文利用相对容易获得的成对约束作为标记数据，将先验信息融入分步的距离计算中，改变了现有方法在框架外加入先验信息的模式，框架内的分步添加模式可以提高算法的效率，同时也提高了成对约束的应用率；其次本文将网络中一些重要节点之间的链接进行断开或者连接，改变了网络结构，使网络结构更加清晰；最后为了保证成对约束的充分利用，本文加入逻辑推理[11]扩大了成对约束的范围，大大提升了成对约束的添加效率。为了避免初始聚类中心随机选取造成的节点归类不稳定，同时保证先验信息的充分利用，用半监督的聚类中心初始化原则，第 2 节中会对初始聚类中心的选取进一步说明。

1.3　节点标注方法

为了弥补半监督学习无法标注最有价值的数据的缺陷，主动学习的策略被运用在社区发现中。通过制定学习策略自动、有效地选出最有指导价值的数据节点，其次由该领域的专家对获取的数据节点进行确认、标志，并加入到标签数据集中。Yang 等人[13]在 2015 年提出了基于非负矩阵分解的主动地半监督社区发现模型，该模型可以自动选择网络中最不稳定的链接，通过计算成员概率分布的信息熵进行标注。Cheng 等人[14]通过网络加权方法找到网络中重要度高的节点进行标注，进而高效利用先验信息。但是信息熵的标注方式，需要得到社区成员对应

的概率分布，这在其他的社区发现算法中是很难得到的，所以这种节点标注方式的选取具有一定局限性。

因此本文提出一种高效通用的节点标注方式，通过计算节点的隶属度来判断节点在本社区的稳定程度。社区中节点的外度 k_{out}，即与该节点相连的节点中不属于本社区的个数，节点的内度 k_{in}，即与该节点相连的节点属于本社区的个数。首先定义粗聚类结果中每个社区的节点隶属度 MD（degree of membership），假设在第一次粗聚类的结果中，节点 i 属于社区在 k，那么节点在其所属社区的隶属度表示为：

$$MD_k(i)=\frac{AN_k(i)}{AN(i)} \tag{2}$$

其中: $AN_k(i)$ 表示社区 k 内所有与节点 i 相连的节点，等同于社区 k 内节点 i 的内度 $z(i)_{in}^k$，$AN(i)$ 表示节点 i 的所有邻接节点（Adjacency Node），等同于节点 i 的度。

$$AN_k(i)=z(i)_{in}^k \tag{3}$$

通过判断一个社区中节点的隶属度来寻找边界节点和中心节点。选取社区中隶属度最高的节点作为中心节点，因为隶属度最高的节点通常是一个社区中最稳定的节点，它的本社区属性最强而且与其他社区的联系最弱。同理，选取社区中隶属度最低的节点作为边界节点。

2 主动地纠错式半监督社区发现算法

本文提出的 ESCD 算法可以主动地选择最不确定的节点，加入少量的标记数据提高社区划分的准确度，具体思路如下：

a）根据已知的先验信息选择初始聚类中心。

b）将 K-means 算法进行分步计算，其中包括：

(a)根据每次的节点距离划分结果，计算当前划分结果社区 k 中每个节点的隶属度 $MD_k(i)$。

(b)根据节点隶属度主动的添加先验信息，并利用成对约束进行逻辑推理，从而改变邻接矩阵结构和聚类索引。

(c)利用改变后的邻接矩阵再次聚类，直到每个节点和其所属社区聚类中心的距离不再变化后停止迭代。

c)得到最终社区划分结果。

下面将对上述步骤里的关键技术进行详细分析。

2.1 初始聚类中心选取原则

K-means 算法依赖于初始聚类中心的选择。大部分研究集中在如何选取聚类中心，Kaufman 等[15]提出了选取数据点局部密度最高作为初始聚类中心的优化方法，Rodriguez 等[16]基于聚类中心比其近邻样本分布密集程度更高，而且与其他密度较高样本距离相对较远的特点，提出了快速搜索密度峰值算法，以密度峰值点作为初始聚类中心；Basu 等人[17]利用标签数据对均值聚类算法进行初始化，提出了两种半监督聚类算法；冷等[18]提出一种新颖的初始中心选择策略，它使用标签数据集辅助选取初始点，保证每个类中至少有一个数据对象被选取。

为了避免初始聚类中心随机选取造成的算法不稳定，采用一种半监督的初始化方案，Gu 等人[19]提出一种半监督的初始中心选择策略，它使用标签数据集辅助初始聚类中心 $\{c_j\}$ 的选取。包含 s 个不同标签的数据集中，取同一类标签几个节点的平均值作为该类的初始聚类中心，总共选取 s 个初始聚类中心。剩余的 K-s 个聚类中心从无标签数据集中选取，选取与已有聚类中心距离最远的节点，这种初始聚类中心的选取方式避免了随机性，保证了所选中心点的分散性。为了保证实验的准确性，在后续的实验中采取十次的聚类结果的 NMI 平均值作为最终的实验结果。

2.2 算法实现

本文提出的主动地纠错式半监督社区检测算法 ESCD，具体步骤如下：

输入：网络 $G=(V,E)$，邻接矩阵 A，社区数 K，收敛条件为节点索引不再改变，迭代次数 $Iter$ 。

输出：社区集 $Setc=\{c_1,\cdots,c_k\}$。

a）按照上述初始聚类中心选取原则选取 K 个节点作为初始聚类中心 $\{c_j\}$，$j=1,2,...,k$；

b）对剩余的每个节点测量其到每个聚类中心的距离 $d(x_i,C_j)$，并把它归到最近中心的一类，并且按照公式（4）更新聚类中心；

$$c_j=\frac{\sum_{i=1}^{l}x_i}{\sum_{i=1}^{l}1}=\frac{\text{类别 } j \text{ 中包含的节点特征和}}{\text{类别 } j \text{ 中的节点个数}} \tag{4}$$

c）根据上一步中的归类结果，找出当前粗聚类结果中每个社区的中心节点{ V_{hub} }和边界节点{ V_{mar} }，根据先验信息中 must-link 和 cannot-link 以及三个规则实现对这些节点之间的边的重构，从而更新网络结构，得到新的邻接矩阵 A^h，$h=1,2,....,Iter$ ；

d）用更新后的邻接矩阵 A^h 重新计算各个节点到聚类中心的距离，并把它归到最近的聚类中心那一类，再次按照式（4）更新聚类中心 $\{c_j\}$；

e）重复 c)d)步直至节点的归类不再变化，算法结束。

2.3 主动地成对约束添加方式

在聚类的过程中发现，被划分错误的那些节点都具有一些共同的特征，尤其是当网络的模块结构比较模糊的时候。导致节点划分错误的主要原因为以下两点：a)社区之间的连接较多，没有明显边界，导致网络结构不够清晰，而且这些连接的端点大部分属于各社区的边界点;b)社区内的一些节点的外度 z_{out} 较大，内度 z_{in} 较小，说明这些节点与本社区外的节点连接较多，与本社区内的节点连接较少。因此只要寻找到这些位于社区边界的节点，就可以通过成对约束来指导这些节点间的边是否连接或断开，通过这样主动学习的方法可以高效地提升社区的清晰度，并且对社区的正确划分有很重要的意义。

根据 2.2 的叙述，基于每次的聚类结果加入少量的标签，将先验信息以成对约束的形式加入其中。将一次完整的 K-means 算法分步计算，基于第一次的距离计算结果加入少量的标记数据，将先验信息以成对约束的形式混入邻接矩阵中，为了使每一个获得的先验信息发挥其最大的作用，采取主动地邻

接节点选择方式来添加成对约束。算法框架图如图 2 所示，这个主动地半监督的过程主要遵循以下三个规则。

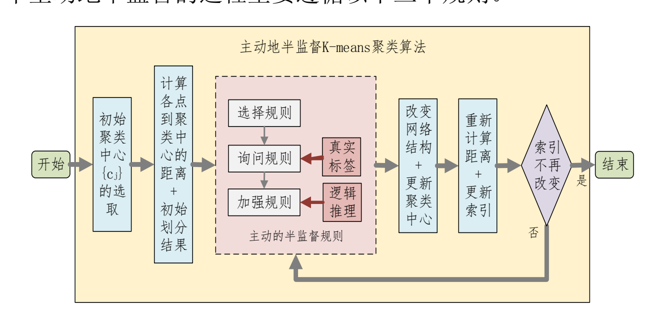


图 2　主动地半监督社区发现算法框架

1）选择规则

根据提出的半监督聚类模型，基于第一次的粗聚类结果，遍历每个社区，选取每个社区的中心节点和一些边界节点，这些边界节点之间的边包含了跨社区连接。

a) 边界节点{ V_{mar} }：这些节点是最不确定且包含最多信息的，所以也包含着部分的跨社区连接。如公式（5）所示，边界节点定义为社区中隶属度最小的节点，如果一个社区中包含多个隶属度值相同的节点则全部标记为边界节点，之后使用先验信息指导这些及节点的划分。

$$V_{mar}(k)=max(MD_k(i)) \tag{5}$$

b) 中心节点{ V_{hub} }：社区中最稳定的节点，如公式（6）所示，中心节点定义为社区中隶属度最大的节点，即与该节点所有相连的节点中位于本社区的比重最大，如果有多个节点都拥有最大的隶属度值，则选择没有与边界点相连的节点作为中心节点，如果都与边界节点相连则随机选择隶属度最高的其中之一。

$$V_{hub}(k)=min(MD_k(i)) \tag{6}$$

因为边界节点的不确定性，所以与边界节点相连的节点真实隶属度可能会改变。其次如图 3-(a)所示是已经选取好中心节点{hub}和边界节点{A，C}，{B}。

2）询问规则

询问规则中，需要用先验信息来指导纠错的操作。遍历每个社区，两两社区之间进行比较，如图 3-(b)所示首先将选择规则中两社区的边界节点{A，B}和{C，B}所有的邻接边进行标记，因为与跨社区的连接相连的边也是不确定的。其次如图 3-(b)所示根据真实标签询问所选择的跨社区连接两端的节点{A，B}和{C，B}的成对约束，如果是正约束则保留这个连接，同时判断两对边界节点所属真实社区标签，如果与 A 相同，则断开所有已经标记的与 B 相连的边；若与 B 相同，反之。经过以上的步骤，如图 3(c)所示已经可以成功地将错误划分到右边社区的 B 节点纠正。

3）加强规则

为了防止询问规则中断开边两端的节点过于稀疏，基于规则 2，根据边界节点和中心节点的真实标签为已经断开边的节点加强连接防止稀疏。如图 3-(d)所示，已知断边的节点 B 和中心节点 V_{hub} 之间存在 must-link，所以将二者连接，同理将右边社区断边的节点进一步加强。加强规则的产生是由于个别社区的节点数较少，连接过于稀疏会导致个别节点在断开连接之后成为孤立节点，易将这些节点错分到其他社区。这样的加强连接可以使社区的块结构更加明显，同时可以避免连边较少的节点成为噪点也防止了规模较小的社区错误划分。此时的社区结构已经较为清晰了，再将更新的聚类中心和重构的网络结构作为下一次 K-means 算法中距离迭代的输入，直到所有节点到聚类中心的距离不再变化停止迭代。

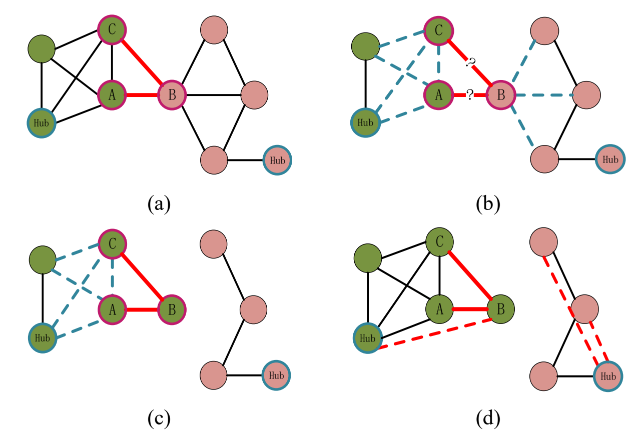


图 3　主动地半监督社区划分过程

3　实验分析

在广泛的人工网络和真实网络中对 ESCD 方法进行了评估。为了检验 ESCD 算法的有效性，与现有的应用最广泛的几种半监督社区发现算法进行了比较，包括 Spin 方法[20]、SNMF 方法[16]、CL-ML 方法[15]。Spin 方法是一种采用节点成对约束的模型，SNMF 是一种对称的非负矩阵分解模型，嵌入了隐含空间的图正则化，CL-ML 是在非负矩阵分解的模型中加入了先验信息的逻辑推理，增强了先验信息的强度。采用广泛的评价标准 NMI 对社区发现的结果进行评价。

3.1　评价标准

根据当前的划分结果 $G_{ESCD}=\{C_1,C_2,...,C_k\}$ 与真实社区 $G=\{C_1^{'},C_2^{'},...,C_K^{'}\}$ 对比来判断社区发现的准确性。其中 G_{ESCD} 是当前的社区划分结果，共有 K 个社区。NMI 是以真实社区结果为标准对当前划分准确度评价指标。NMI（归一化互信息）表达公式如下：

$$NMI=\frac{-2\sum_{ij}N_{ij}\log\frac{N_{ij}N_t}{N_{i.}N_{.j}}}{\sum_i N_{i.}\log\frac{N_{i.}}{N_t}+\sum_i N_{.j}\log\frac{N_{.j}}{N_t}} \tag{7}$$

3.2　实验设计

本文实验设计主要分为两个部分，分别在真实网络和生成网络中对提出的纠错式半监督聚类模型进行评估。6 个真实网络的节点数 N 和社区数 K 如表 1 所示。

表 1　真实网络数据集

数据集	Karate	dolphins	football	School6	School7	polbooks
N	34	69	115	69	69	105
K	2	2	12	6	7	3

相比以上六个真实网络数据，生成网络模拟了真实网络中节点度和社区大小的无标度性质，在聚类研究中经常使用。生成网络的常见参数如下：网络中的节点数目 N；网络中节点的平均度值 k；最小社区所拥有的节点数 minc；最大社区所拥有的节点数 maxc；混合参数 μ，表示节点与其他社区的节点的连接概率，μ 取值越大，说明社区结构越复杂，聚类分析越困难。如表 2 所示，在这里采用 K_{out}=7 和 K_{out}=8 的 GN 网络以及 μ=0.65 和 μ=0.75 的 LFR 网络进行实验。

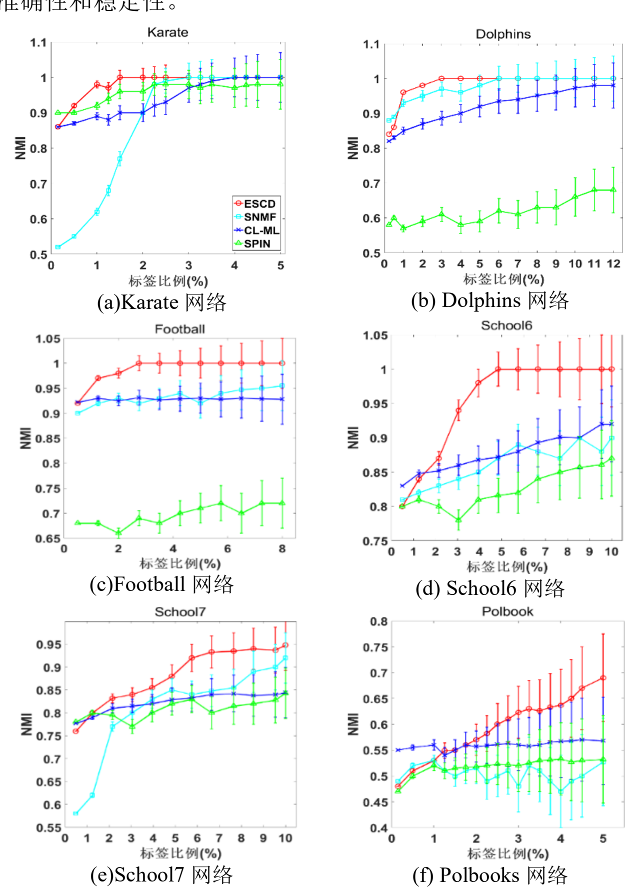
表 2　生成网络数据集参数设置

数据集	N	K	K_{out}	μ	minc	Maxc
GN-7	128	4	7	/	32	32
GN-8	128	4	8	/	32	32
LFR-0.65	1000	29	/	0.65	20	50
LFR-0.75	1000	29	/	0.75	20	50

3.3　实验结果

3.3.1 真实网络实验结果

如图 4 所示，在六个广泛的真实社区对本文方法进行评估，图中的四条曲线分别代表提出的算法与对比试验的结果，为了测试实验的稳定性，对每个方法进行十次实验，其中曲线上的节点表示十次实验结果的平均值，而纵向的直线表示十次结果的方差。从实验结果来看，提出的方法加入少量标签就可以达到很好的性能，例如在 Footballs 社区中，只需要加入 2.32%的标签，NMI 的值就可以从 0.921 提升到 1。六个真实网络的实验结果都可以验证提出方法的有效性和稳定性，通过在分步距离计算中主动地加入成对约束，可以大大提升 K-means 算法的准确性和稳定性。



(a)Karate 网络　(b) Dolphins 网络

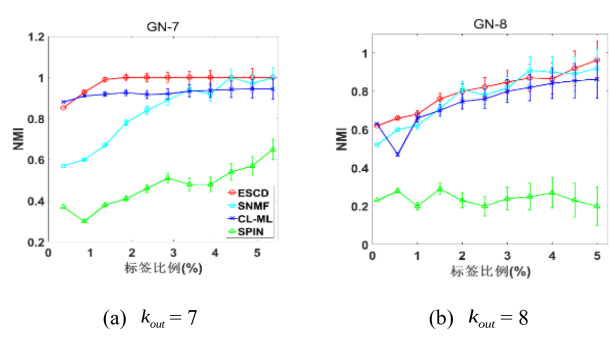
(c)Football 网络　(d) School6 网络

(e)School7 网络　(f) Polbooks 网络

图 4　真实网络社区发现实验结果

3.3.2 生成网络实验结果

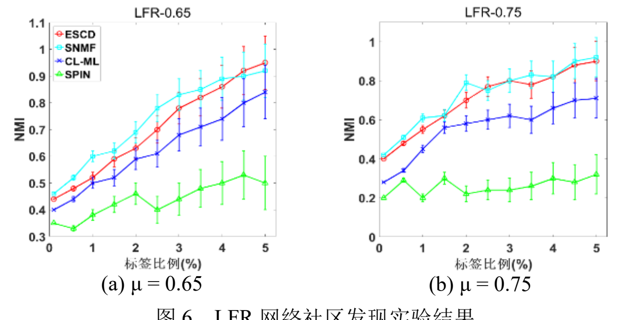
为了验证提出的 ESCD 方法在生成网络中的有效性，首先在 GN 网络进行实验，图 5 中是十次实验的平均结果。横坐标加入成对约束的比例，纵坐标表示 NMI 值，达到 1 表示社区划分结果与实际情况完全一致。与其他对比方法相比，加入标签比例相同时，提出的 ESCD 方法划分更加准确。在 GN 网络 K_{out}=7 时，ESCD 方法加入 2%的标签就可以将性能提升至 1。当 K_{out}=8 时，ESCD 方法加入 1%和 2%的标签时 NMI 值分别可以达到 0.713 和 0.807，性能优于其他的半监督算法。



(a)　$k_{out}=7$　　(b)　$k_{out}=8$

图 5　GN 网络社区发现实验结果

为了进一步验证本文方法的有效性，在两个 LFR 网络中进行测试。如图 6(a)所示，提出的方法和 SNMF 表现都很好，可以在标签量很少的时候达到理想的结果，在加入 5%左右的标签时本文方法超过了 SNMF。



(a) μ = 0.65　　(b) μ = 0.75

图 6　LFR 网络社区发现实验结果

3.4　划分过程演示

本节将以真实社区为例呈现主动地纠错式半监督划分过程，根据实验结果观察通过邻接选择方式主动的添加先验信息，实际上是基于当前粗划分结果对划分错误的节点不断地修正。在 Football 中主要展现其中几个社区之间错误划分如何纠正，图 7 是 Football 在第一次距离计算的索引结果，有 6 个社区是完全划分正确的，主要的划分错误集中在了中间的几个社区，为了更直观的了解算法的计算过程，截取其中的一部分来对本文算法详细说明。

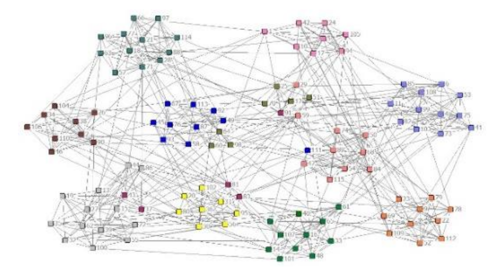


图 7　Football 社区粗聚类结果

首先根据当前粗聚类结果，两两社区依次遍历。图 8 所示的两个主要社区有多个节点交错，在真实情况中同一颜色的节点属于同一社区。每两个社区之间的判断都会进行三次，尽可能多的判断不确定节点。如图 8 所示的划分过程，首先根据选择规则选取出这两个社区的中心节点和边界节点。根据计算公式得到两个社区的中心节点分别是 29 和 50，为了更直观地观察实验过程，在图 8 将中心节点放大作为标记。两个社区的边界点分别是 59 和 74,89（74 与 89 拥有相同的隶属度值）。

接下来根据真实标签询问节点 59 和 74,89 是否属于同一社区，（‘等号’表示两个节点拥有 must-link），即{59=74}，并且它们之间没有连接，在他们之间加一条红色的边；{59=89}属于同一社区，他们之间存在连接，所以保持连接，图中用绿色的表示保持连接。根据询问规则，应该断开 59 与其所在社区的所有边，但是由于 59 与它所在社区内的其余节点并无连接，所以无须断开边，这也更加说明，社区的边界点是与本社区连接最不紧密的节点。

最后根据加强规则，接下来判断节点 74 和 89 与中心节点 50 的标签是否一致，通过之前的两次查看真实标签，得到了{59=74}，{59=89}根据逻辑判断可以得知{74=89=59}。接下来与中心节点 50 的连接关系只需要确定这三个节点其中之一即可得到加强的结果。例如只判断节点 74 和 50，根据真实标签，二者属于同一社区，所以再一次利用逻辑判断将 59，74，89 都与 50 连接。在上述的一次纠错过程中，通过使用聚类结果和逻辑判断节省了大量的标签信息，本文算法可以通过最少的标签获得最多的判断结果。

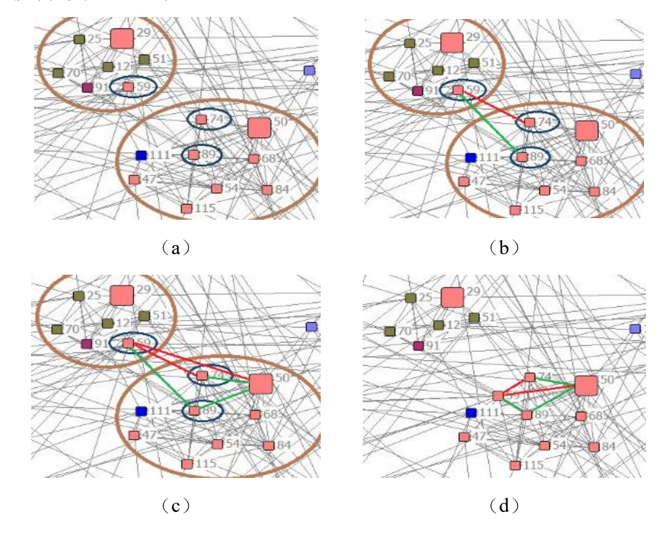


图 8　主动纠错过程 1

59 号节点通过半监督纠错算法已经被划分到了正确的社区，接下来通过相同的规则，进行下两个社区的纠错，如图 9(a)所示左侧蓝色社区中有三个节点被划分错误，第一次节点 64，60，70 分别被选为两个社区的边界节点，通过半监督规则，最终将 64，60 与 70 之间连接，并且断开与左侧社区的连边，此时观察到右边社区中心节点的真实标签与本社区并不相同，所以这里提醒了与中心节点的连接也需要成对约束的指导，但是这并不意味着需要很多标签，由于本文的约束对包含逻辑推理所以只需判断其中一个节点与中心节点的标签即可。如图 9(c)所示，第二次将 98 和 70 选为两个社区的边界节点，最终节点 60，64，98 都被正确划分到右边的社区。两个社区之间的主动纠错工作就已经完成了，接下来遍历其他社区，在这里不再一一说明。

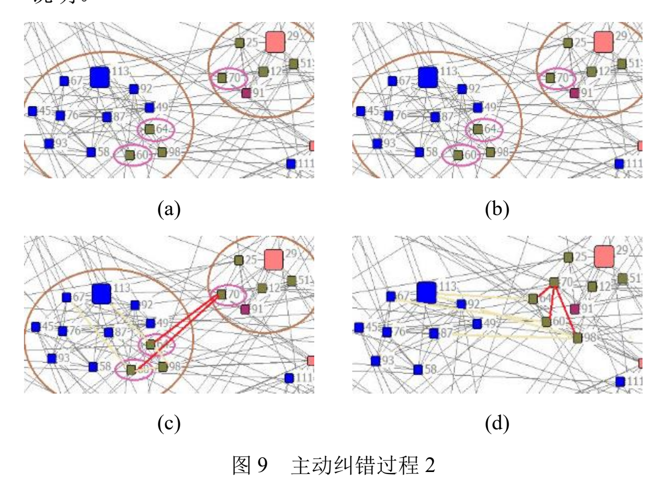


图 9　主动纠错过程 2

4　结束语

本文在分析了现有聚类算法的特点和不足的基础上，提出的 ESCD 算法在原有的聚类算法中分步计算距离，每一步中主动地选择不确定的节点加入先验信息，通过三个规则实现整理复杂网路的连接关系，使得网络的模块性大大提升。通过理论分析在真实数据集上的实验表明，ESCD 算法能够显著提高社区发现的准确性和稳定性。

在接下来的工作中，将进一步考虑将该半监督模型应用在社交网络，比如微博等，而且引入更多的参考因素社区发现，节点重要性度量方法和种子节点的选择方式，比较引入多个参考因素的算法对真实社区划分的影响，并考虑在算法中加入深度学习对复杂的社区结构进行特征提取，使算法具有更高的实用价值。

参考文献：

[1] Lin Youfang, Wang Tianyu, Tang Rui. An effective model and algorithm for community detection in social networks [J]. Journal of Computer Research and Development, 2012, 49 (2): 337-345

[2] Newman M E. The Structure and Function of Complex Networks [J]. SIAM Review, 2003, 45 (2): 168-256.

[3] Perozzi B, Akoglu L. Focused clustering and outlier detection in large attributed graphs [C]// Proc of ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2014: 1346-1355.

[4] MacQueen J. Some methods for classification and analysis of multivariate observations [C]// Proc of the 5th Berkeley Symposium on Mathematical Statistics and Probability. 1967: 281-297.

[5] Allahverdyan A E, Steeg G V, Galstyan A. Community detection with and without prior information [J]. Europhysics Letter, 2009, 90 (1): 983-995.

[6] Ma Xiaoke, Gao Lin, Rong Xue, *et al.* Semi-supervised clustering algorithm

for community structure detection in complex networks [J]. Physica A Statistical Mechanics & Its Applications, 2012, 389 (1): 187-197.

[7] Leng Mingwei, Yao Yukai, Cheng Jianjun, *et al.* Active semi-supervised community detection algorithm with label propagation [C]// Proc of International Conference on Database Systems for Advanced Applications. Berlin: Springer, 2013: 324-338.

[8] Zhu Xiaojin. Semi-supervised learning with graphs [D]. Pittsburgh, PA: Carnegie Mellon University, 2005.

[9] Liu Dong, Bai Hongyu, Li Huijia, *et al.* Semi-supervised community detection using label propagation [J]. International Journal of Modern Physics B, 2014, 28 (29): 145-208.

[10] Silva T C, Liang Zhao. Semi-supervised learning guided by the modularity measure in complex networks [J]. Neuro Computing. 2012, 78 (1): 30-37.

[11] Zhang Zhongyuan, Sun Kaidi, Wang Siqi. Community structure detection in complex networks with partial background information [J]. Scientific Reports, 2013, 3 (11): 32-41.

[12] Yang Liang, Cao Xiaochun, Jin Di, *et al.* A unified semi-supervised community detection framework using latent space graph regularization [J]. IEEE Trans on Cybernetics, 2017, 45 (11): 2585-2598.

[13] Yang Liang, Jin Di, Wang Xiao, *et al.* Active link selection for efficient semi-supervised community detection [J]. Scientific Reports, 2015, 5: 9039.

[14] Cheng Jianjun, Leng Mingwei, Li Longjie, *et al.* Active semi-supervised community detection based on must-link and cannot-link constraints [J]. PLOS One, 2014, 9 (10): 88-110.

[15] Kaufman L, Rousseeuw P J. Finding groups in data: an introduction to cluster analysis [J]. The Astrophysical Journal, 1998. 508 (1) .

[16] Rodriguez A, Laio A. Machine learning Clustering by fast search and find of density peaks [J]. Science, 2014, 344 (6191): 1492-1496.

[17] Basu S, Banerjee A, Mooney R J. Semi-supervised clustering by seeding [C]// Proc of the 19th International Conference on Machine Learning. 2002, 27-34.

[18] Leng Mingwei, Huang Liang, Li Longjie, *et al.* Active semi-supervised community detection based on asymmetric similarity measure [J]. International Journal of Modern Physics B, 2015, 29 (13): 1550078-.

[19] Gu Lei, Lu Xianling. Semi-supervised subtractive clustering by seeding [C]// Proc of International Conference on Fuzzy Systems & Knowledge Discovery. 2012: 738-741.

[20] Eaton E, Mansbach R. A spin-glass model for semi-supervised community detection [C]// Proc of the 26th AAAI Conference on Artificial Intelligence. 2012: 900-906.